How to measure bacterial genome plasticity? A novel time-integrated index helps gather insights on pathogens

Greta Bellinzona[1], Gherard Batisti Biffignandi[1], Fausto Baldanti[2,3], Matteo Brilli[4], Davide Sassera[1,5], Stefano Gaiarsa[3]*.

1. Department of Biology and Biotechnology. University of Pavia. Pavia, Italy

2. Department of Medical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia. Pavia, Italy

3. Microbiology and Virology Unit. Fondazione IRCCS Policlinico San Matteo. Pavia, Italy

4. Department of Biosciences. Pediatric Clinical Research Center Romeo ed Enrica Invernizzi. University of Milan. Milan, Italy

5. Fondazione IRCCS Policlinico San Matteo. Pavia, Italy

*corresponding

Address correspondence to: s.gaiarsa@smatteo.pv.it

Abstract (207/250)

Genome plasticity can be defined as the capacity of a bacterial population to swiftly gain or lose genes. The time factor plays a fundamental role for the evolutionary success of microbes, particularly when considering pathogens and their tendency to gain antimicrobial resistance factors under the pressure of the extensive use of antibiotics. Multiple metrics have been proposed to provide insights into the gene content repertoire, yet they overlook the temporal component, which has a critical role in determining the adaptation and survival of a bacterial strain. In this study, we introduce a novel index that incorporates the time dimension to assess the rate at which bacteria exchange genes, thus fitting the definition of plasticity. Opposite to available indices, our method also takes into account the possibility of contiguous genes being transferred together in one single event. We applied our novel index to measure plasticity in three widely studied bacterial species: *Klebsiella pneumoniae*, *Staphylococcus aureus*, and *Escherichia coli*. Our results highlight distinctive plasticity patterns in specific sequence types and clusters, suggesting a possible correlation between heightened genome plasticity and globally recognized high-risk clones. Our approach holds promise as an index for predicting the emergence of strains of potential clinical concern, possibly allowing for timely and more effective interventions.

Introduction

The evolutionary success of an organism hinges on its capacity to continuously adapt to changes encountered within the environment. Prokaryotes are notably characterized by the ability to acquire exogenous DNA through horizontal gene transfer (HGT) (Ochman et al. 2000) and easily lose unnecessary genes (Morris et al. 2012), both representing powerful adaptation tools. The balance between gene gain and loss events constitutes a complex trade-off (Brockhurst et al. 2019): while the acquisition of genes enables the emergence of novel functions, it also may impose a metabolic burden on the organism, demanding energy and resources for the synthesis and maintenance of the acquired genes. Whether a gene is kept, is determined by the equilibrium between the fitness advantages within the specific environment and the metabolic cost associated. For instance, the presence of antibiotic resistance genes introduces a trade-off whereby a resistant strain will outcompete susceptible bacteria in the presence of antibiotics, but the metabolic burden in the absence of antibiotics will disadvantage the resistant strain (Andersson and Levin 1999; Andersson and Hughes 2017; Rajer and Sandegren 2022).

Not only the presence of such mechanisms is fundamental for the evolution of a bacterial population, but the time component plays a key role in determining their effectiveness. The capacity of a bacterial community to swiftly gain or lose genes, termed genome plasticity, is an important factor contributing to its evolutionary success.

In the era of massive genomics data, investigating genome plasticity on both large and short time scales has become possible. To do so, however, *ad-hoc* tools must be designed. Over the years, several metrics have been proposed to assess genome plasticity and offer insights into the dynamics of genetic exchange among bacteria, such as Jaccard distance applied to gene content (Hernández-González et al. 2018; Wyres et al. 2019). Jaccard distance, a dissimilarity measure widely used in machine learning and computational genomics (Besta et al. 2020), can be applied to compare gene content across different sets of genomes using the following formula:

$$JD = \frac{\sum_{A,B=1\ldots N}^{A<B} \frac{U_A + U_B}{M_A + M_B - S_{A,B}}}{Np}$$

where U_A and U_B are the number of genes found only in genomes a and b respectively and M_A and M_B are the total number of genes found in a and b respectively, $S_{A,B}$ are the shared genes between a and b, and N_p the total number of pairs considered.

In this context, Jaccard distance allows to compare the overall gene repertoire across different groups of genomes. Moreover, Jaccard distance applied to gene content resembles the genome fluidity formula (Kislyuk et al. 2011):

$$Fluidity = \frac{\sum_{A,B=1...N}^{A<B} \frac{U_A + U_B}{M_A + M_B}}{Np}$$

where U_A and U_B are the number of gene families found only in genomes A and B respectively, M_A and M_B are the total number of gene families found in a and b respectively, and N_p the total of genome pairs.

Although the difference is slight, when Jaccard distance is used to compare gene content between two genomes, it assigns lower importance to core genes than genome fluidity. This is achieved by preventing the double counting of shared gene families between the two genomes. Genome fluidity was proposed by the authors as an alternative to core and pan genome for measuring gene content diversity within a species or groups of closely related organisms.

Neither Jaccard distance nor genome fluidity take into account the time component, which as previously stated, is a key component in defining genome plasticity. As a consequence, by applying these two indexes to gene content it is not possible to distinguish between a scenario where a bacterial community gradually accumulates a large gene repertoire over an extended period and a second scenario where the population undergoes rapid gain or loss events. The latter is a distinctive feature of emerging high risk clones in many bacterial pathogen species, one commonly associated with increased virulence, enhanced transmissibility, and extensive antibiotic resistance (Woodford et al. 2011; Del Barrio-Tofiño et al. 2020; Peirano et al. 2020). Additionally, both indexes consider genes as single entities,

whereas HGT events often involve regions with more than one gene. Consequently, even with a time correction, the number of gene gain or loss events may be overestimated.

In this study we aimed at providing a new metric to assess the rate at which bacteria exchange genes. We implemented a new method to compute the number of gene gain and loss events and introduced the evolutionary distance into the calculation. Then, we used the new index to investigate the dynamics of bacterial genome plasticity within three widespread, widely studied bacterial species: *Klebsiella pneumoniae*, *Staphylococcus aureus*, and *Escherichia coli*.

Results and discussion

A new index: Flux of Genes Segments

To assess genome plasticity, we introduced a novel index called Flux of Genes Segments (FOGS). This new index aligns with our definition of genome plasticity, representing the rate of gene acquisition or loss events. For a group of genomes FOGS is defined as:

$$FOGS = \frac{\sum_{A,B=1\ldots N}^{A<B} \frac{N.\ of\ gene\ gain\ or\ loss\ events}{d_{A,B}}}{N_p}$$

d is the SNP distance between genome *A* and genome *B*, *Np* is the total number of genome pairs considered (*Np* = 2/[N(N-1)], where N is the number of genomes considered). The higher the value of FOGS, the higher is the genome plasticity.

Following the most parsimonious hypothesis, we consider the gain or loss of neighboring related genes as a single event rather than distinct ones. Our approach to calculate the number of events is presented in Figure 1. We represented each genome into a graph connecting consecutive genes, akin to the method employed in the computation of the Genome Organization Stability index (Brilli et al. 2013). When two genomes are compared, shared genes are removed, and only unique genes are considered. This avoids considering internal rearrangements. Exploiting tables of gene coordinates, if genes are consecutive they will be compressed and considered as single entities. By summing the total number of unique gene stretches, we get the putative number of gene gain or loss events between the two genomes (Figure 1). Technical details are provided in the Materials and Methods section. Using this strategy, a gain or loss event of consecutive genes is considered as a single event rather than distinct events (Figure 1). It should be pointed out that, using this approach, it is not possible to distinguish whether an event is a gene gain in one genome or a loss in the other one. Moreover, by considering only the unique connections, internal rearrangements do not have an impact on the index.

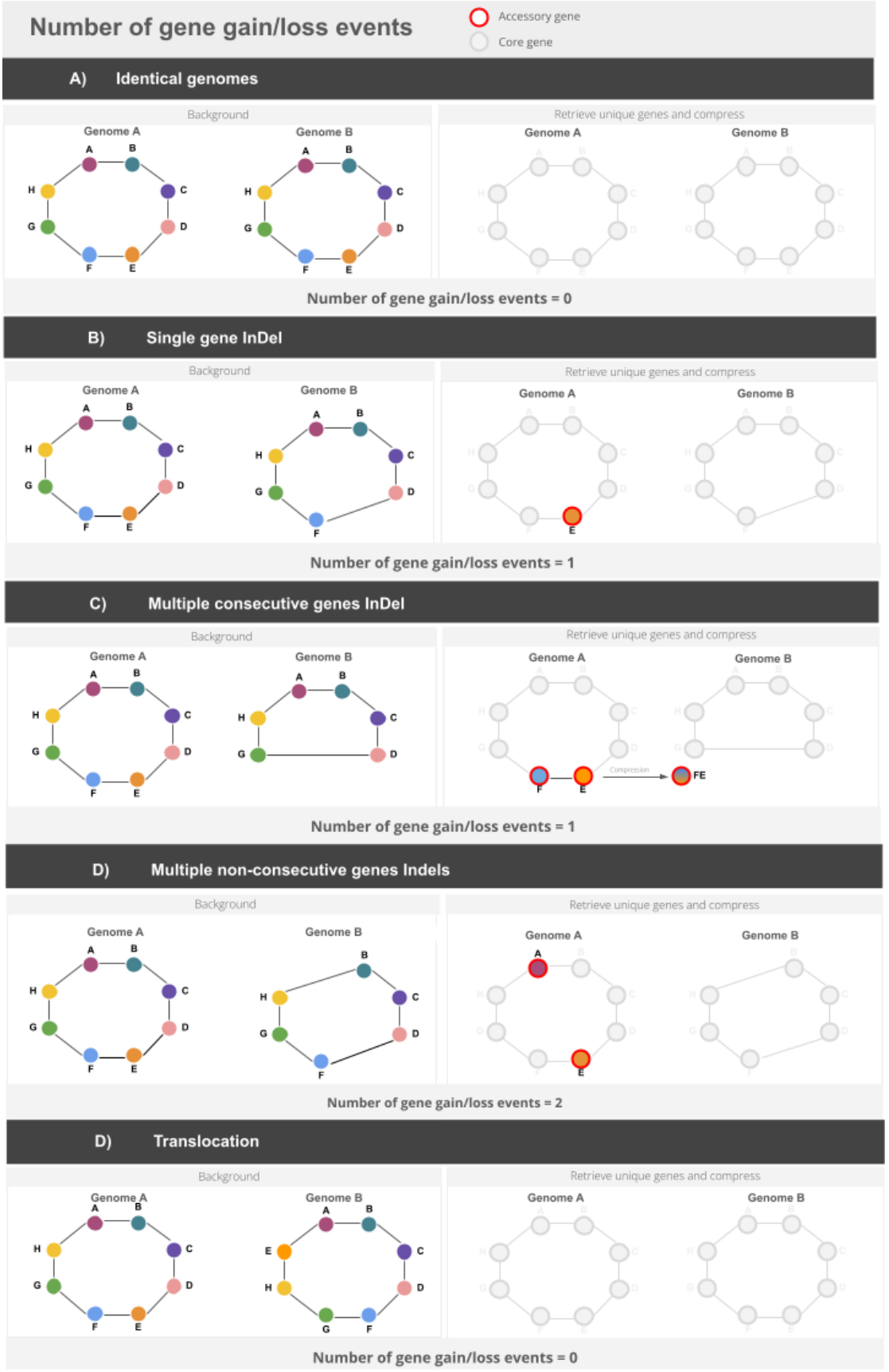

Figure 1. Strategy applied to compute the total number of genes gain/loss events considering different scenarios: A) Identical genomes, B) Single gene Indels, C) Multiple consecutive gene Indel, D) Multiple non-consecutive gene indels and E) Translocations.

Once the number of putative genes gain or loss events have been determined for each pairwise comparison, the temporal scaling is introduced, dividing the obtained value by the evolutionary distance. To measure evolutionary distance we decided to use the core SNP distance. This choice was made because our aim was to apply the index to closely related strains. Core SNPs are particularly useful for evolutionary purposes since they preserve a higher discriminant power compared, for instance, to single copy orthologs (Bush et al. 2020)

Application of plasticity indexes to large genome datasets

We applied FOGS to investigate plasticity trends within three species of clinical interest. We choose as case studies *K. pneumoniae*, which is a well known nosocomial bacterium often associated with multidrug resistance, *S. aureus,* which is both nosocomial and community acquired, and *E. coli*, known for its wide repertoire of ecological niches. For each organism, we identified prominent taxonomic clusters using the fastBAPS algorithm (Tonkin-Hill et al. 2019) from a comprehensive collection of high-quality genomes accessible through the BV-BRC database (Olson et al. 2023). We then selected a representative subset of 100 genomes from each of the most prominent clusters (>100 genomes). For convenience of interpretation, we labeled each cluster with the most represented Multi-Locus Sequence Type (MLST or ST) among the genomes. See Table 1 for an overview of the results obtained at each stage of the filtering process (for technical details see the Materials and Methods section).

Species	HQ genomes	MLST*	dRep$^{\Delta}$	N. of clusters$^{\&}$	N. of genomes$^{\#}$
Klebsiella pneumoniae	17887	17435	11196	19	1900
Staphylococcus aureus	18849	18188	11319	20	2000
Escherichia coli	42049	40805 (#1, Atchman)	22099	35	3500

Table 1. Datasets building. * Number of genomes for which it was possible to obtain the ST. ∆ Number of genomes after the dereplication step. & Clusters detected by fastBAPS comprised at least 100 genomes. # Number of genomes used for subsequent analysis.

We applied the newly created index FOGS to the three species datasets in order to investigate the presence of different plasticity patterns among the clones. Initially, we considered the presence of a relationship between the number of gene gain or loss events versus the number of SNP for each pairwise comparison (Supplementary Figure 1) within each species. Although a general relationship was observed—indicating that a higher SNP distance corresponds to a higher number of potential gene gain/loss events—there is a noticeable clustering at low SNP distances. This clustering reveals a broad spectrum of putative events in this range. Also, the relationship observed is non-linear.

In addition, we investigate a possible association between FOGS and virulence or resistance traits by comparing the index to the mean number of resistance and virulence genes for each cluster. As follows, we discuss the results obtained for each species in depth.

Klebsiella pneumoniae

K. pneumoniae is one of the leading causes of hospital-acquired infections, especially among immunocompromised individuals, elderly patients, and those with chronic conditions. Within the hospital environment, this bacterium can persist on various surfaces, medical equipment, and in water sources. Thus, it poses a significant challenge in terms of antibiotic resistance, making it a persistent concern for healthcare settings (Anon 2022). The *K. pneumoniae* population shows a remarkable level of diversity, encompassing numerous distinct phylogenetic lineages or 'clones' (Holt et al. 2015). These lineages can be defined as Multidrug Resistant (MDR), or hypervirulent based on the prevalence of determinants of resistance to antibiotics or virulence genes.

Wyres and colleagues analyzed the gene content diversity between MDR and hypervirulent clones and found a reduced genetic diversity in the hypervirulent ones. They achieved this result by calculating the pairwise Jaccard gene content distances among genomes belonging to a clone (Wyres et al. 2019). In our study, we focused on evaluating the rate at which genome content evolves. Most of the clusters in our dataset, including STs 681, 13, 218, 348, 231, 395, 65, and 29 displayed a generally low gene flow. However, a smaller subset of

clusters, exclusively represented by STs 307, 101, and 15 exhibited higher values, suggesting greater plasticity (Figure 2).

Clusters at higher FOGS levels (307, 101, 15, 147) correspond to STs that are widely recognized globally as emerging high-risk clones, due to their potential to cause severe infections and their association with antimicrobial resistance (AMR), including pan resistance (Heiden et al. 2020; Loconsole et al. 2020; Peirano et al. 2020; Arcari and Carattoli 2022). Notably, ST307 has been gradually displacing the well known ST258 in multiple areas of the world (e.g.(Peirano et al. 2020)). Readers should take into account that the cluster we named '258' encompasses ST11, ST258 and ST512 which are part of the CG258. This lineage has been for a long time the most prevalent MDR-Carbapenemase producer group (Munoz-Price et al. 2013; Arcari and Carattoli 2022).

Based on our analysis, cluster 307 exhibits a substantially higher genome plasticity compared to cluster 258 ($p < 1E-10$). This finding provides a possible explanation for the recent rise of ST307, which appears to be the most plastic. Moreover, considering the mean number of resistance genes per clone, ST307 stands out as one of the most resistant clusters in our dataset. A positive correlation between plasticity and antibiotic resistance can be observed (Supplementary Figure 2A, R^2=0.44, p-value<0.05). This result is in accordance with the findings of Wyres and coworkers (Wyres et al. 2019). However, our results also suggest that no direct or inverse relationship is present between virulence and plasticity (R^2=0.14, p-value>0.05) (Supplementary Figure 2B). To hypothesize an explanation to these observations, one should consider that virulence genes mostly imply a constant fitness advantage. So they are more likely to be maintained and transmitted vertically than resistance genes, which are only needed in the presence of the antimicrobial and repeatedly lost and regained (e.g.(Simner et al. 2018)). As a consequence, our results suggest that the *K. pneumoniae* strains with a higher plasticity are also more likely to be resistant to antimicrobials.

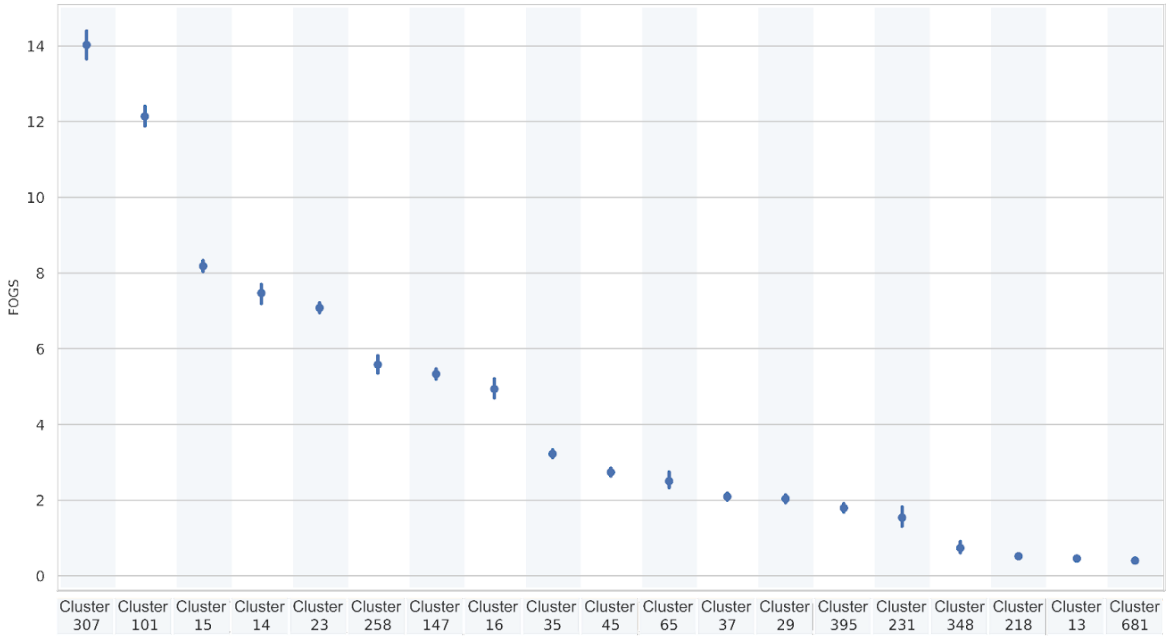

Figure 2. Pointplot FOGS within each cluster identified by fastBAPS (Tonkin-Hill et al. 2019) in *K. pneumoniae*. Vertical lines represent the confidence intervals.

Staphylococcus aureus

S. aureus is a versatile Gram-positive bacterium that colonizes the skin and mucous membranes of humans and animals. While it is a common member of the human microbiota, *S. aureus* can also cause a wide range of infections, from minor skin and soft tissue infections to life-threatening diseases such as bloodstream infections, pneumonia, and endocarditis (Tabah and Laupland 2022). As *K. pneumoniae*, *S. aureus* is known for its ability to acquire and maintain resistance to multiple antimicrobial agents, making it a significant public health concern worldwide (Howden et al. 2023). Among resistant strains, methicillin-resistant *Staphylococcus aureus* (MRSA) is particularly challenging due to the limited availability of alternative treatment options (Kourtis et al. 2019). Genetic factors such as *mec* genes are responsible for this resistance (Vandendriessche et al. 2013) .

ST5 and ST8 are the two major STs in *S. aureus* and are commonly associated with various types of infections, including those acquired in healthcare settings as well as in the community (Monaco et al. 2017; Strauß et al. 2017). In our study, the fastBAPS algorithm successfully divided both ST5 and ST8 into distinct subgroups. We observed a noteworthy

difference in plasticity within the subgroups of ST8, specifically the cluster 8A which exhibited significantly higher plasticity compared to the other subgroup, cluster 8B ($p < 1E-10$) (Figure 3). We hypothesized the presence of a highly successful and adaptable sub-strain within ST8, which exhibits the highest level of plasticity in our entire dataset. Subsequently, we investigated the presence of resistance and virulence determinants in these subgroups. Notably, cluster 8A was found to be enriched in specific resistance genes: *mecA* (responsible for methicillin resistance) was found in all the 8A genomes but only in the 60% of the 8B group (Supplementary Figure 3); *mph(C)* and *mrs(A)* confer resistance to macrolide antibiotics, such as erythromycin, azithromycin, and clarithromycin; *aph(3')-III* is associated to the resistance to gentamicin, tobramycin, and amikacin; *ant(6)-Ia* provides resistance to aminoglycoside antibiotics, including kanamycin and neomycin. While the virulence pattern remained relatively stable between the two subgroups, we discovered a significant difference in the presence of CHIPS genes, which were found in the majority of genomes within cluster 8A. These genes play a crucial role as important virulence factors, helping *S. aureus* evade the innate immune defense systems (van Wamel et al. 2006). These findings underscore the significance of cluster 8A as a particular plastic subgroup within the ST8. On the other hand, the two subgroups of ST5 did not show a significant difference in plasticity, resistance or virulence patterns.

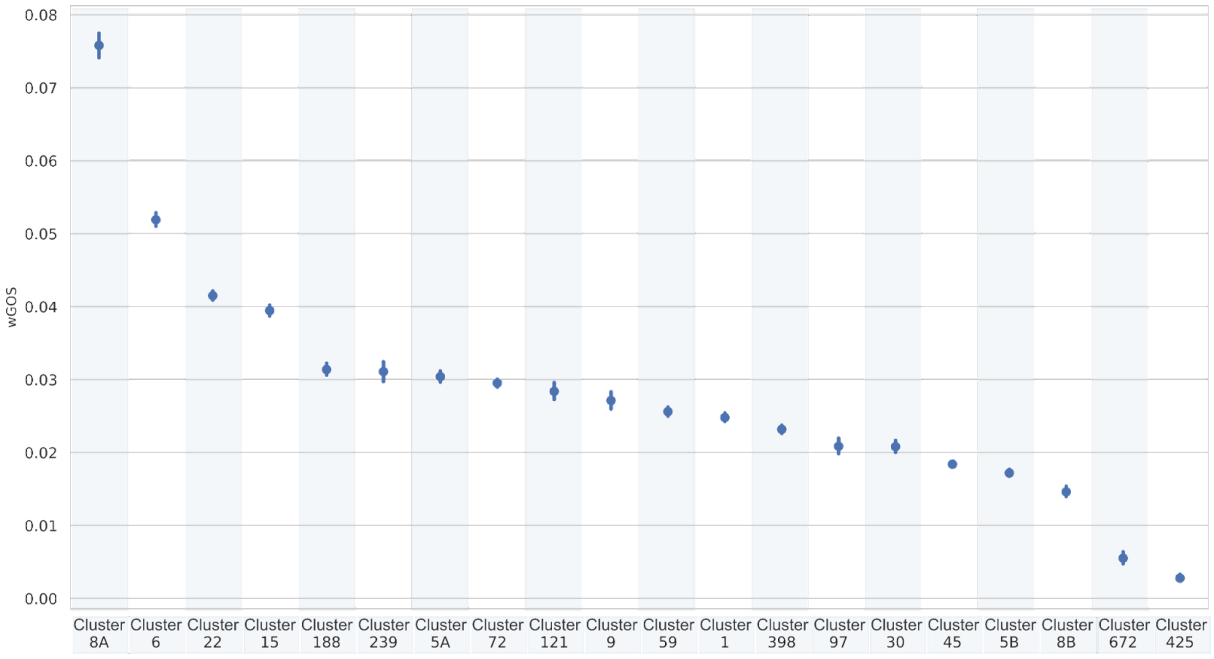

Figure 3. Pointplot FOGS within each cluster identified by fastBAPS (Tonkin-Hill et al. 2019) in *S. aureus*. Vertical lines represent the confidence intervals.

Escherichia coli

E. coli is a widely studied Gram-negative bacterium that plays a crucial role in various ecological niches and has significant implications for human health. Previous studies have demonstrated the presence of extensive genetic variation within *E. coli*, enhanced by its ability to adapt to a range of diverse niches including hospitals, animal reservoirs, and natural ecosystems (Blount 2015; Jang et al. 2017; Koh et al. 2022). In the clinical setting, *E. coli* represents a significant public health concern due to its ability to cause a wide range of infections, ranging from urinary tract infections to more severe bloodstream infections.

ST131 is one of the most predominant and globally disseminated lineages associated with urinary tract infections (UTIs) and bloodstream infections. Frequently associated with multidrug resistance, including extended-spectrum beta-lactamases (ESBLs) and fluoroquinolone resistance, it can be considered the most successful MDR clone of all time (Cummins et al. 2021; Peirano et al. 2022). In our dataset, ST131 was split in two clusters, namely '131A' and '131B'. Cluster '131A' showed the highest FOGS value of the dataset, and significantly higher in respect to cluster '131B' ($p<0.01$; Figure 4). ST131's population structure has been widely investigated and three genetically distinct clades have been identified (A, B and C), each characterized by different fimbrial adhesin (*fimH*) gene variants

(Decano and Downing 2019). Our analysis revealed that cluster 131B uniquely contains genomes encoding the *fimH*41 variant, which is exclusive to the globally recognized cluster A. In contrast, cluster 131A encompasses various *fimH* variants, including 30 and 40, thereby comprising the global clusters B and C. Within specific subclades of the global cluster C, a convergence of extensive resistance and virulence profiles has been observed (Biggel et al. 2022). Biggel and colleagues propose that this convergence may not be applicable to other *E. coli* lineages, such as ST73 and ST95. Despite being pandemic, these lineages exhibit low antibiotic resistance, potentially attributed to their gene acquisition capabilities (Biggel et al. 2022). Our findings align with this hypothesis, revealing a lower degree of genomic plasticity in ST73 and ST95.

Interestingly, cluster 1193 appears to be the second most plastic clone according to FOGS. *E. coli* ST1193 is currently emerging rapidly across the globe, mimicking the very successful ST131 (Pitout et al. 2022).

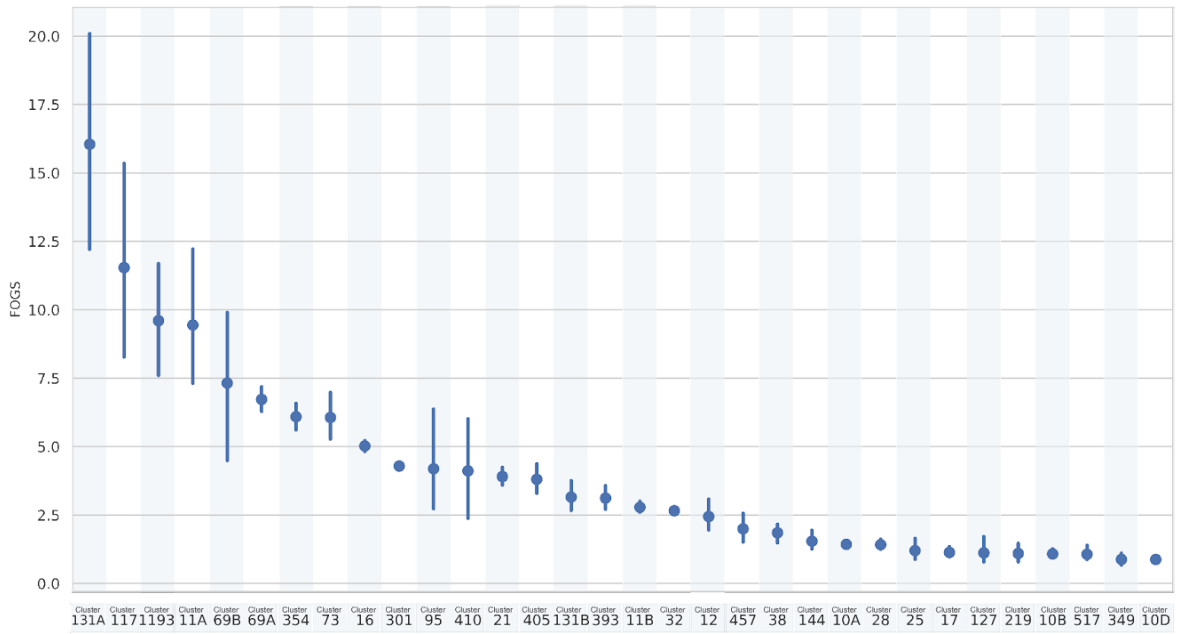

Figure 4. Pointplot FOGS within each cluster identified by fastBAPS (Tonkin-Hill et al. 2019) in *E. coli*. Vertical lines represent the confidence intervals.

Conclusions

The balance between gene gain and loss events represents a complex trade-off for prokaryotes, shaping their adaptive strategies in response to environmental challenges (Brockhurst et al. 2019).

Here, we introduced a novel index, Flux of Genes Segments (FOGS), to assess genome plasticity by quantifying the rate of gene acquisition or loss events in a group of genomes. The integration of the temporal component in FOGS aligns with the key role of time in defining the effectiveness of adaptive mechanisms within bacterial communities.

We applied FOGS to analyze genome plasticity trends in three bacterial species: *K. pneumoniae*, *S. aureus*, and *E. coli*. Our findings reveal distinct plasticity patterns within specific sequence types and clusters, with a possible connection to globally recognized high-risk clones. Consequently, FOGS may be used to recognize and predict emerging strains of clinical importance, even at the subclone/subST level.

Materials and methods

Datasets construction

A curated dataset of high quality genomes was collected from BV-BRC (Wattam et al. 2017) (updated December 2022) independently for *K. pneumoniae*, *S. aureus* and *E. coli* using the makepdordb.py script of the P-DOR pipeline (Batisti Biffignandi et al. 2023), which automatically filters for genome size and number of contigs. To further improve the quality of each dataset, we performed *in silico* Multilocus Sequence Typing (MLST), using schemes downloaded from PubMLST (Jolley et al. 2018) in December 2022. For *E. coli* we chose the Achtman scheme. Since MLST is based on single-copy housekeeping genes, the absence, or the presence of more than one copy, of one of these genes is most likely owing to a poor quality genome assembly. Only genomes that could be assigned a Sequence Type (ST), using an *in-house* python script, were used in the next phase.

To reduce redundancy (e.g. to avoid almost identical genomes that were obtained to analyze outbreaks) from each dataset, dRep (Olm et al. 2017) was applied using the "dereplicate" function (-ms 10000 -pa 0.99 --SkipSecondary), also allowing to confirm the quality of the genome using its internal default pipeline.

A SNP alignment was obtained for the reduced datasets using P-DOR (-n 0) (Batisti Biffignandi et al. 2023) and then used as input for fastBAPS (Tonkin-Hill et al. 2019) with default settings. After eliminating any fastBAPS-assessed clusters with less than 100 genomes, a random selection of 100 genomes was chosen from each cluster that was still present to ensure representativeness. We named each cluster with the prevalent ST; when a ST was split between clusters, we added a letter to the name.

SNP distances

A SNP alignment was generated on the final datasets using a P-DOR (-n 0) (Batisti Biffignandi et al. 2023) with an internal complete genome as reference for each species (*E. coli*: CP043539; *S. aureus*: LT963437; *K. pneumoniae*: CP006648). Pairwise SNP distances were computed using the snp-dists tool (https://github.com/tseemann/snp-dists).

Flux of Genes Segments

In order to compute FOGS, we followed the same approach of Brilli and colleagues (Brilli et al. 2013) by translating each genome into a graph. (1) Orthofinder (Emms and Kelly 2019) was used to classify all the proteins from each species dataset into orthology groups. (2) The gene neighborhood network of each genome was built using the information about coding genes location in the annotation file previously produced by prokka (Seemann 2014). Each gene was connected to the one downstream in the genome table only if their distance was less than 1000 bp and if they belonged to the same contig. We then performed the pairwise graphs (genomes) comparison as follows: (3) for each genome the unique genes are retrieved, (4) and compressed if consecutive, considering them as a single element. (5) The sum of the unique gene segments obtained corresponds to the number of gene gain or loss events. To this purpose, we used the connected_components function included in the SciPy python library (Virtanen et al. 2020). (6) The number of gene gain or loss events was then weighted on the SNP distance, computed as previously described. All the steps were performed using an *in-house* python script.

Resistance and Virulence genes

To assess the presence of resistance genes, we utilized Resfinder (Florensa et al. 2022) (Florensa et al., 2022). Our criteria for determining gene presence required a minimum of 60% query coverage and 90% sequence identity. Similarly, we assessed the presence of virulence genes using the Virulence Finder Database (VFDB) (Chen et al. 2005). Analyses on *K. pneumoniae* were repeated using Kleborate (Lam et al. 2021).

To ascertain the presence and variant of *fimH*, a dedicated BLASTn search was performed using the fimtyper database (bitbucket.org/genomicepidemiology/fimtyper_db) as reference (100% query coverage, 100% sequence identity).

Statistical analysis

The mean distribution of wFOGs values between fastBAPS clusters were tested using the Kruskal-Wallis test, followed by Dunn's test for pairwise comparisons using Benjamini-Hochberg adjustment. The analyses were performed using R v4.1.3.

Data availability

The scripts used to compute FOGS are available at https://github.com/MIDIfactory/Genome-Plasticity

Funding

This research was supported by EU funding within the NextGenerationEU-MUR PNRR Extended Partnership initiative on Emerging Infectious Diseases (Project no. P E00000007, INF-ACT)

Supplementary Material Legends

Supplementary Figure 1. Density plots of Number of gene gain or loss events against the SNP distance for all pairs within each fastBAPS cluster for K. pneumoniae, S. aureus and E. coli.

Supplementary Figure 2. A) Mean number of resistance genes against FOGS within each K. pneumoniae cluster. The regression line is presented in gray (R2=0.44 , p-value<0.05). B) Mean number of virulence genes against FOGS within each K. pneumoniae cluster. The regression line is presented in gray (R2=0.14 , p-value>0.05).

Supplementary Figure 3. Relationship between FOGS and the mean number of resistance genes in Klebsiella pneumoniae. The color of the point indicates the prevalence of the mecA within each cluster.

Bibliography

Andersson DI, Hughes D. 2017. Selection and Transmission of Antibiotic-Resistant Bacteria. *Microbiol Spectr* [Internet] 5. Available from: http://dx.doi.org/10.1128/microbiolspec.MTBP-0013-2016

Andersson DI, Levin BR. 1999. The biological cost of antibiotic resistance. *Curr. Opin. Microbiol.* 2:489–493.

Anon. 2022. Klebsiella species: Taxonomy, hypervirulence and multidrug resistance. *eBioMedicine* 79:103998.

Arcari G, Carattoli A. 2022. Global spread and evolutionary convergence of multidrug-resistant and hypervirulent Klebsiella pneumoniae high-risk clones. *Pathog. Glob. Health* [Internet]. Available from: https://www.tandfonline.com/doi/abs/10.1080/20477724.2022.2121362

Batisti Biffignandi G, Bellinzona G, Petazzoni G, Sassera D, Zuccotti GV, Bandi C, Baldanti F, Comandatore F, Gaiarsa S. 2023. P-DOR, an easy-to-use pipeline to reconstruct bacterial outbreaks using genomics. *Bioinformatics* [Internet] 39. Available from: http://dx.doi.org/10.1093/bioinformatics/btad571

Besta M, Kanakagiri R, Mustafa H, Karasikov M, Ratsch G, Hoefler T, Solomonik E. 2020. Communication-efficient jaccard similarity for high-performance distributed genome comparisons. In: 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS). IEEE. Available from: https://ieeexplore.ieee.org/document/9139876/

Biggel M, Moons P, Nguyen MN, Goossens H, Van Puyvelde S. 2022. Convergence of virulence and antimicrobial resistance in increasingly prevalent Escherichia coli ST131 papGII+ sublineages. *Commun Biol* 5:752.

Blount ZD. 2015. The unexhausted potential of E. coli. *Elife* [Internet] 4. Available from: http://dx.doi.org/10.7554/eLife.05826

Brilli M, Liò P, Lacroix V, Sagot M-F. 2013. Short and long-term genome stability analysis of prokaryotic genomes. *BMC Genomics* 14:309.

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The Ecology and Evolution of Pangenomes. *Curr. Biol.* 29:R1094–R1103.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 33:D325–D328.

Cummins EA, Snaith AE, McNally A, Hall RJ. 2021. The role of potentiating mutations in the evolution of pandemic Escherichia coli clones. *Eur. J. Clin. Microbiol. Infect. Dis.* [Internet]. Available from: http://dx.doi.org/10.1007/s10096-021-04359-3

Decano AG, Downing T. 2019. An Escherichia coli ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.* 9:17394.

Del Barrio-Tofiño E, López-Causapé C, Oliver A. 2020. Pseudomonas aeruginosa epidemic high-risk clones and their association with horizontally-acquired β-lactamases: 2020 update. *Int. J. Antimicrob. Agents* 56:106196.

Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. 2022. ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom* [Internet] 8. Available from: http://dx.doi.org/10.1099/mgen.0.000748

Heiden SE, Hübner N-O, Bohnert JA, Heidecke C-D, Kramer A, Balau V, Gierer W, Schaefer S, Eckmanns T, Gatermann S, et al. 2020. A Klebsiella pneumoniae ST307 outbreak clone from Germany demonstrates features of extensive drug resistance, hypermucoviscosity, and enhanced iron acquisition. *Genome Med.* 12:113.

Hernández-González IL, Moreno-Hagelsieb G, Olmedo-Álvarez G. 2018. Environmentally-driven gene content convergence and the Bacillus phylogeny. *BMC Evol. Biol.* 18:148.

Holt KE, Wertheim H, Zadoks RN, Baker S, Whitehouse CA, Dance D, Jenney A, Connor TR, Hsu LY, Severin J, et al. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae* , an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.* 112:E3574–E3581.

Howden BP, Giulieri SG, Wong Fok Lung T, Baines SL, Sharkey LK, Lee JYH, Hachani A, Monk IR, Stinear TP. 2023. Staphylococcus aureus host interactions and adaptation. *Nat. Rev. Microbiol.* 21:380–395.

Jang J, Hur H-G, Sadowsky MJ, Byappanahalli MN, Yan T, Ishii S. 2017. Environmental Escherichia coli: ecology and public health implications-a review. *J. Appl. Microbiol.* 123:570–581.

Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 3:124.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12:32.

Koh XP, Shen Z, Woo CF, Yu Y, Lun HI, Cheung SW, Kwan JKC, Lau SCK. 2022. Genetic and Ecological Diversity of and Cryptic Clades in Subtropical Aquatic Environments. *Front. Microbiol.* 13:811755.

Kourtis AP, Hatfield K, Baggs J, Mu Y, See I, Epson E, Nadle J, Kainer MA, Dumyati G, Petit S, et al. 2019. Vital Signs: Epidemiology and Recent Trends in Methicillin-Resistant and in Methicillin-Susceptible Staphylococcus aureus Bloodstream Infections - United States. *MMWR Morb. Mortal. Wkly. Rep.* 68:214–219.

Lam MMC, Wick RR, Watts SC, Cerdeira LT, Wyres KL, Holt KE. 2021. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nat. Commun.* 12:4188.

Loconsole D, Accogli M, De Robertis AL, Capozzi L, Bianco A, Morea A, Mallamaci R, Quarto M, Parisi A, Chironna M. 2020. Emerging high-risk ST101 and ST307 carbapenem-resistant Klebsiella pneumoniae clones from bloodstream infections in Southern Italy. *Ann. Clin. Microbiol. Antimicrob.* 19:24.

Monaco M, Pimentel de Araujo F, Cruciani M, Coccia EM, Pantosti A. 2017. Worldwide Epidemiology and Antibiotic Resistance of Staphylococcus aureus. *Curr. Top. Microbiol. Immunol.* 409:21–56.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* [Internet] 3. Available from: http://dx.doi.org/10.1128/mBio.00036-12

Munoz-Price LS, Poirel L, Bonomo RA, Schwaber MJ, Daikos GL, Cormican M, Cornaglia G, Garau J, Gniadkowski M, Hayden MK, et al. 2013. Clinical epidemiology of the global expansion of Klebsiella pneumoniae carbapenemases. *Lancet Infect. Dis.* 13:785–796.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.

Olm MR, Brown CT, Brooks B, Banfield JF. 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11:2864–2868.

Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, et al. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res.* 51:D678–D689.

Peirano G, Chen L, Kreiswirth BN, Pitout JDD. 2020. Emerging Antimicrobial-Resistant High-Risk Klebsiella pneumoniae Clones ST307 and ST147. *Antimicrob. Agents Chemother.* [Internet] 64. Available from: http://dx.doi.org/10.1128/AAC.01148-20

Peirano G, Chen L, Nobrega D, Finn TJ, Kreiswirth BN, DeVinney R, Pitout JDD. 2022. Genomic Epidemiology of Global Carbapenemase-Producing Escherichia coli, 2015-2017. *Emerg. Infect. Dis.* 28:924–931.

Pitout JDD, Peirano G, Chen L, DeVinney R, Matsumura Y. 2022. Escherichia coli ST1193: Following in the Footsteps of E. coli ST131. *Antimicrob. Agents Chemother.* 66:e0051122.

Rajer F, Sandegren L. 2022. The Role of Antibiotic Resistance Genes in the Fitness Cost of Multiresistance Plasmids. *MBio* 13:e0355221.

Simner PJ, Antar AAR, Hao S, Gurtowski J, Tamma PD, Rock C, Opene BNA, Tekle T, Carroll KC, Schatz MC, et al. 2018. Antibiotic pressure on the acquisition and loss of antibiotic resistance genes in Klebsiella pneumoniae. *J. Antimicrob. Chemother.* 73:1796–1803.

Strauß L, Stegger M, Akpaka PE, Alabi A, Breurec S, Coombs G, Egyir B, Larsen AR, Laurent F,

Monecke S, et al. 2017. Origin, evolution, and global transmission of community-acquired ST8. *Proc. Natl. Acad. Sci. U. S. A.* 114:E10596–E10604.

Tabah A, Laupland KB. 2022. Update on Staphylococcus aureus bacteraemia. *Curr. Opin. Crit. Care* 28:495–504.

Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. 2019. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* 47:5539–5549.

Vandendriessche S, Vanderhaeghen W, Soares FV, Hallin M, Catry B, Hermans K, Butaye P, Haesebrouck F, Struelens MJ, Denis O. 2013. Prevalence, risk factors and genetic diversity of methicillin-resistant Staphylococcus aureus carried by humans and animals across livestock production sectors. *J. Antimicrob. Chemother.* 68:1510–1516.

van Wamel WJB, Rooijakkers SHM, Ruyken M, van Kessel KPM, van Strijp JAG. 2006. The innate immune modulators staphylococcal complement inhibitor and chemotaxis inhibitory protein of Staphylococcus aureus are located on beta-hemolysin-converting bacteriophages. *J. Bacteriol.* 188:1310–1315.

Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, et al. 2017. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* 45:D535–D542.

Woodford N, Turton JF, Livermore DM. 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* 35:736–755.

Wyres KL, Wick RR, Judd LM, Froumine R, Tokolyi A, Gorrie CL, Lam MMC, Duchêne S, Jenney A, Holt KE. 2019. Distinct evolutionary dynamics of horizontal gene transfer in drug resistant and virulent clones of Klebsiella pneumoniae. *PLoS Genet.* 15:e1008114.